Siamese Instance Search for Tracking - Supplementary Material

Ran Tao, Efstratios Gavves, Arnold W.M. Smeulders
QUVA Lab, Science Park 904, 1098 XH Amsterdam
{r.tao,egavves,a.w.m.smeulders}@uva.nl

1. Illustrations of network architectures

To learn the matching function that operates on pairs of data, we use a Siamese architecture with two branches [1, 2]. The Siamese network processes the two inputs separately through individual networks that take the form of a convolutional neural network. For individual branches, we investigate two different network architectures, a small one adapted from AlexNet [5] (Figure 1a) and a very deep one inspired by VGGNet [7] (Figure 1b).

2. Additional Sequences

The strengths of the proposed Siamese INstance search Tracker (SINT) are further illustrated on 6 newly collected sequences from YouTube. The sequences have considerable degrees of scale change, fast motion, low contrast, out-of-plane rotation, illumination variation, non-rigid deformation and poorly textured objects. The videos showing the sequences together with the tracking results of SINT and two recent trackers, MEEM [8] and MUSTer [4], are available on YouTube with the following URLs: "Fishing"(https://youtu.be/K-70sLC6gRU), "Rally"(https://youtu.be/QiCDDQTGcn4), "BirdAttack"(https://youtu.be/r3SgEuuUhDY), "Soccer"(https://youtu.be/1GYzl79iXtk), "GD"(https://youtu.be/gWWHmSCgSno), "Dancing"(https://youtu.be/oMG1pJZSno0). The tracking results of SINT are highlighted in red, while the results of MEEM and MUSTer are in green and blue respectively. The annotations are available every 5 frames, shown in yellow.

3. Tracking Target Re-identification

We observe that provided with a candidate sampling over the whole frame using [9], the proposed SINT tracker has good capability of re-identifying the target after the target was missing for a significant amount of time from the video. This is illustrated on a 1500-frame, 12-shot *Star Wars* video where we track *Yoda* (https://youtu.be/knaxUljyY_Q). We show that SINT is able to discover *Yoda* when it re-enters the camera view after being absent for a complete shot. We also indicate the absence of the target in this video by simply thresholding the matching scores with the original observation in the first frame with a threshold 0.65 in this illustration. We will investigate a more principled way of declaring presence or absence in the future work.

References

[1] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah. Signature verification using a siamese time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):669–688, 1993. 1

[2] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005. 1

[3] R. Girshick. Fast r-cnn. *arXiv preprint arXiv:1504.08083*, 2015. 2

[4] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao. Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking. In *CVPR*, 2015. 1

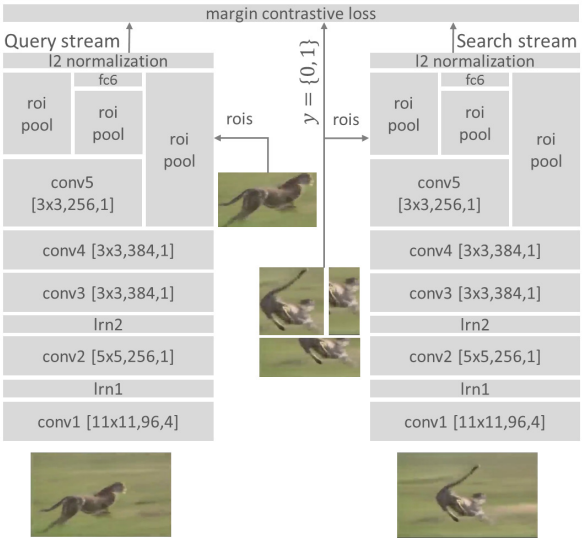
[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[6] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010. 2

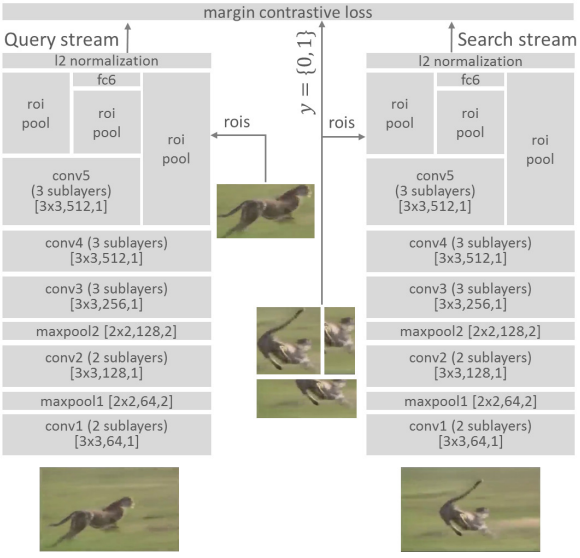
[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 1

[8] J. Zhang, S. Ma, and S. Sclaroff. Meem: Robust tracking via multiple experts using entropy minimization. In *ECCV*, 2014. 1

[9] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 1



(a)



(b)

Figure 1: The proposed two-stream Siamese networks to learn the generic matching function for tracking. 'conv', 'lrn', 'maxpool', 'roipool' and 'fc' stand for convolution, local response normalization, max pooling, region-of-interest pooling [3] and fully connected layers respectively. Numbers in square brackets are kernel size, number of outputs and stride. The fully connected layer has 4096 units. All conv layers are followed by rectified linear units (ReLU) [6].